Validation of the European French Version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V_f)

[*]Timothy Pommée, [†]Margaux Shanks, [‡]Dominique Morsomme, [†]Sandrine Michel, and [*]Ingrid Verduyckt,
[*]*Québec, Canada,* †*Toulouse, France, and* ‡*Liège, Belgium*

Abstract: Objective. This study aimed to validate the French adaptation of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V_f) for assessing voice disorders in France. The CAPE-V_f addresses limitations of the GRBAS by providing a more sensitive, standardized approach to evaluating six vocal parameters (overall severity, roughness, breathiness, strain, pitch, and loudness) on three tasks (sustained vowels, sentence reading, and spontaneous speech). The study focused on investigating the intra- and inter-rater reliability, as well as the convergent and discriminant validity of the CAPE-V_f.

Methods. Thirty-four dysphonic and seven euphonic native French speakers participated in the study. Thirteen speech-language pathologists from France evaluated the voice samples using both the CAPE-V_f and GRBAS tools at a one-week interval. Intra- and inter-rater reliability were calculated using intraclass correlation coefficients (ICC), while convergent and discriminant validity were measured by correlating CAPE-V_f with GRBAS and Voice Handicap Index (VHI) scores, respectively.

Results. The CAPE-V_f showed good intra-rater reliability for overall severity (mean ICC: 0.89), strain (ICC: 0. 83), and pitch (ICC: 0.88), while roughness, breathiness, and loudness exhibited moderate reliability. Inter-rater reliability was low for most parameters, except overall severity, which demonstrated good reliability (mean ICC: 0.77). Strong correlations were observed between CAPE-V_f and GRBAS Grade (mean *r*: 0.84), supporting its convergent validity. Moderate correlations were found for roughness, breathiness, and strain. The CAPE-V_f's correlation with the VHI was moderate (mean *r*: 0.53), reflecting its discriminant validity.

Conclusion. The CAPE-V_f is a valid and reliable tool for perceptual assessment of voice disorders in French-speaking populations, with stronger psychometric properties than the GRBAS, particularly for intra-rater reliability and overall severity. While inter-rater reliability was lower, qualitative feedback suggested that improvements to the protocol, particularly for pitch and loudness ratings, could enhance its clinical applicability. The findings support the CAPE-V_f as a comprehensive tool for standardized clinical voice assessment.

Key Words: Voice assessment—Auditory-perceptual—CAPE-V—French—Validity—Reliability..

INTRODUCTION

The European Laryngeal Society's guideline for the multidimensional assessment of voice disorders requires a combination of (a) a videolaryngostroboscopic investigation by a ear, nose, and throat physician (ENT); (b) instrumental measures; (c) a perceptual voice quality assessment by the clinician; and (d) a self-assessment of voice-related quality of life by the patient.[1,2] To date, perceptual rating of the patient's voice remains the gold standard in clinical voice assessment.[3] This method is considered the most clinically relevant and ecologically valid tool for analyzing the voice, given that voice itself is inherently a perceptual phenomenon that is meant to be heard.[4] The most common tools used for perceptual assessment[5] are the GRBAS[6] and the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V).[7]

The GRBAS is composed of five simple four-point scales and is widely used by clinicians and researchers[8] to assess the overall grade of severity (G), roughness (R), breathiness (B), astheny (A), and strain (S). The GRBAS can be used with any vocal production, such as a sustained vowel or spontaneous speech for a more natural representation of the patient's voice.[9] Although long used as a reference tool, the GRBAS has been the subject of various studies questioning its validity and reliability.[8,10–12] Two major limitations of this tool are (1) the scoring on Likert scales and (2) the absence of a standardized protocol, which decreases the intra- and inter-rater reliability of the ratings.[8]

To address these limitations, the CAPE-V was developed during a consensus conference by the ASHA "Special Interest Group 3—Voice and Voice Disorders" in June 2002 in Pittsburgh, USA. In the original American version, perceptual assessment is conducted based on a standardized protocol consisting of two sustained vowels, six sentences eliciting specific vocal behaviors (eg, soft and hard glottal attacks, nasality), and a spontaneous speech sample in response to the question "Tell me about your voice problem" or "Tell me how your voice is functioning". Six voice characteristics are assessed on a 100-mm hybrid visual analog scale (VAS) labeled with categorical markers

Accepted for publication October 17, 2024.

From the [*]École d'orthophonie et d'audiologie, Faculté de médecine, Université de Montréal, Québec, Canada; †Département de Médecine, maïeutique et paramédical, Faculté de Santé, Université Toulouse III Paul Sabatier, Toulouse, France; and the ‡Research Unit for a Life-Course Perspective on Health and Education, Université de Liège, Liège, Belgium.

Address correspondence and reprint requests to TimothyPommée, 7077 avenue du Parc, local 3001–1, Montréal, Québec, H3N 1X7. E-mail: timothy_pommee@hotmail.com

Journal of Voice, Vol xx, No xx, pp. xxx–xxx
0892-1997

https://doi.org/10.1016/j.jvoice.2024.10.021

(mild, moderate, and severe): overall severity, roughness, breathiness, strain, pitch, and loudness. If the vocal impairment varies by task, the evaluator places multiple markers and numbers them accordingly (eg, #1 for the sustained vowel task, #2(a) for the first sentence, etc). To the right of each scale, the clinician can further indicate if the vocal quality is consistently or intermittently present. Two blank VAS are also provided for the clinician to assess other vocal characteristics if needed. Finally, a space is provided for additional comments on resonance and other voice characteristics such as diplophonia, fry, and falsetto. The CAPE-V's standardized administration protocol allows for a more comprehensive assessment of vocal behaviors and enhances the rating reliability. Additionally, the use of VAS allows for a more sensitive and precise scoring, providing data that are more suitable for statistical analysis[13] and achieving better reliability than Likert scales.[8,11] Various studies have confirmed its psychometric superiority (eg, better inter-rater reliability[8,14] and higher sensitivity to the fine parameters of voice disorders[12]).

Given the influence of sociocultural factors—including language both on the speaker's and on the rater's side—on the perception and description of voice quality,[15–21] the CAPE-V has been adapted and validated in more than 12 languages.[14,22–38] As no standard guidelines exist for these adaptation efforts, a heterogeneity of methods and expert collaborations was employed, as highlighted in a recent systematic review.[38] In 2023, a French adaptation of the CAPE-V (CAPE-V_f) was developed in Belgium, as part of an international collaboration between voice researchers in Belgium, France, and Quebec.[39] After consulting a linguist for initial translation of the CAPE-V sentences considering the originally targeted vocal behaviors and the specificities of the French language, this adaptation effort used a three-round Delphi process. An expert panel was iteratively consulted about the task stimuli (vowels, sentences, and question), the rating scales, and the wordings to describe the evaluated vocal characteristics. The final protocol of the CAPE-V_f, consensually approved by the Delphi expert panel after the third round, consists of the sustained vowel /a/, six sentences, and an emotionally neutral question to elicit semispontaneous speech. A significant modification was made to the scoring system of the CAPE-V_f: while the American version and adaptations in other languages include the labels "MI" ("mildly deviant"), "MO" ("moderately deviant"), and "SE" ("severely deviant") beneath each VAS, the creators of the French version decided to remove these markers with the consensus of experts during the Delphi study. This decision aimed to avoid the clustering of ratings around the verbal severity markers, as also highlighted by Nagle.[40]

Given the superior psychometric properties of the CAPE-V compared with the GRBAS as established by previous studies, the present study aimed to answer the following question: Is the French adaptation of the CAPE-V a reliable and valid tool for the perceptual assessment of voice in France? Three hypotheses were investigated: (1) the CAPE-V_f demonstrates good intra- and inter-rater reliability (ICC ≥ 0.75) for all assessed parameters and its reliability is higher than that of the GRBAS; (2) the CAPE-V_f demonstrates good convergent validity, as measured by a strong correlation ($r \geq 0.70$) with the different parameters of the GRBAS; (3) the correlation between the overall severity score of the CAPE-V_f and the total Voice Handicap Index (VHI) score is weaker than its correlation with the "G" score of the GRBAS ($r < 0.70$), given that these two tools evaluate different constructs of the multidimensional vocal phenomenon (discriminant validity[41]).

METHODS

The first phase of the study involved collecting voice recordings from dysphonic and euphonic individuals. The second phase involved the assessment of these voices by experienced speech-language pathologists (SLPs) using both the CAPE-V_f and the GRBAS, at a 1-week interval.

Participants

Speakers

Thirty-four dysphonic speakers (Table 1) were recruited among the patients of four SLPs who specialized in voice therapy in private practices located in the Occitanie and Bourgogne-Franche-Comté regions of France. The inclusion criteria were as follows: adult patients, diagnosed with dysphonia, and native French speakers. Seven euphonic subjects (Table 1) were recruited via word-to-mouth by the second author MS. The only requirement for participation was the absence of a history of voice disorders. All participants received an information sheet along with consent forms for data processing.

SLPs—raters

Approximately 950 SLPs listed in the Ostéovox training registry[42] were contacted by email, as well as ENT specialists and phoniatricians through their professional email addresses. Fifty-five professionals, including 54 SLPs and one phoniatrician, initially showed interest in participating. The inclusion criteria for the final sample of raters were as follows: SLP, phoniatrician, or ENT specialist with at least one year of experience in the field of voice. Twenty-one SLPs confirmed their participation in the study. They were divided across four listening lists. However, due to time constraints, only thirteen SLPs shared the results of their assessment, thereby modifying the composition of the pre-established listening lists (list 1: two raters, lists 2 and 3: three raters, and list 4: five raters). The final sample of raters consisted of seven SLPs with 1-10 years of experience in the field of voice, two with 11-20 years of experience, and three with 21-30 years of experience. One SLP had more than 30 years of experience. The raters were from five different regions across France (Pays de la Loire, Île-de-France, Grand Est, Auvergne-Rhône-Alpes, and Occitanie).

TABLE 1.
Demographic Data for the Speakers

	Dysphonic		Euphonic	
Sex	Male (*n* = 13)	Female (*n* = 21)	Male (*n* = 4)	Female (*n* = 3)
Age *mean (SD), [range]*	54.46 (18.68), [23-83]	56.76 (13.09), [30-90]	37.25 (15.80), [25-57]	46.33 (20.31), [23-60]
Diagnosis (*n*)				
Functional dysphonia	3	5	-	-
Recurrent laryngeal paralysis	2	3	-	-
Singing voice disorder	1	3	-	-
Nodules	0	2	-	-
Presbyphonia	1	1	-	-
Scarring	1	1	-	-
Postoperative cyst	0	2	-	-
Postoperative polyp	1	0	-	-
Oropharyngeal cancer	1	0	-	-
Chemotherapy	0	1	-	-
Laryngeal papillomatosis	0	1	-	-
Laryngeal dyskinesia	1	0	-	-
Edema	1	0	-	-
Puberphonia	1	0	-	-
Stroke	0	1	-	-
Undetermined	1	0	-	-

Procedure

Voice samples

Voice samples were recorded using a standardized procedure. A recording protocol was provided to the clinicians, specifying the required equipment, and recording conditions, the procedure for informing and obtaining consent from patients, the method for pseudonymization and data transfer, and details about the vocal tasks. All recordings were securely transmitted via FileSender Renater, hosted by the University of Toulouse. To familiarize clinicians with the recording procedure and verify recording quality before data collection, each clinician performed a trial recording using their own voice.

Speakers were recorded in a quiet and distraction-free environment, using a headset or standing microphone positioned at a 45° angle and 6-10 cm from the mouth. Recordings were made in mono mode, with a resolution of at least 16 bits and a signal sampling rate of at least 22 kHz. To prevent vocal fatigue or vocal warmup effects in dysphonic subjects, recordings were done before the voice therapy session, after completing the French VHI questionnaire.[43,44]

The speakers first sustained the vowel /a/ for 3-5 seconds, three times. They then read the six sentences presented on flashcards, one at a time. Finally, about 20 seconds of semispontaneous speech was prompted with the question: "Briefly introduce yourself as Jean/Jeanne Dupont, mentioning your region of origin and your main activity." Speakers were free to make up details to avoid sharing personal information. The recordings were listened to and sorted by the second author MS, to ensure good sound quality (clear sound, absence of disruptive background noise) and adherence to the CAPE-V$_f$ protocol. Minor adjustments were made to some recordings, such as increasing the volume.

Listening lists

The retained recordings were categorized into groups based on the consensus-perceived severity of the voice impairment (Table 2) by the three authors TP, MS and SM: no impairment, mild, moderate, and severe impairment. These voices were then divided into four listening lists to ensure a fair distribution of severity levels and sex (Table 2). For feasibility reasons, the number of recordings per list was limited to nineteen, to avoid fatigue and reduced attention. Within each list, seven recordings were repeated to allow for the assessment of intra-rater reliability.

Assessment protocols

Two assessment protocols were created, one starting with the CAPE-V$_f$ (for raters of lists 1 and 2) and one starting with the GRBAS (for raters of lists 3 and 4). This counterbalancing aimed to minimize potential bias related to the order of use of the two tools, at a minimum one-week interval.[14] An exclusively digital format of the CAPE-V$_f$ was used, as a fillable PDF to be completed on a computer. To adapt the CAPE-V$_f$ to this format, the instruction to measure the distance on the VAS scale as well as the space

TABLE 2.
Number of Voices for Each Severity Grade and Sex, and Distribution of Voices Across the Four Listening Lists

Speakers	Severity	*n*	List 1	List 2	List 3	List 4
Men	NO	6	2 + 1rep	1	1	2 + 1rep
	MI	3	1 + 1rep	2 + 1rep	1 + 1rep	2 + 1rep
	MO	5	2 + 1rep	2 + 1rep	1 + 1rep	1 + 1rep
	SE	3	2 + 1rep	1 + 1rep	2 + 1rep	1 + 1rep
Women	NO	6	1	2 + 1rep	2 + 1rep	1
	MI	6	2 + 1rep	1 + 1rep	2 + 1rep	1 + 1rep
	MO	6	1 + 1rep	1 + 1rep	2 + 1rep	2 + 1rep
	SE	6	1 + 1rep	2 + 1rep	1 + 1rep	2 + 1rep
Total		41	12 + 7rep *n* = 19	12 + 7rep *n* = 19	12 + 7rep *n* = 19	12 + 7rep *n* = 19

Abbreviations: rep, repeated recordings; NO, no voice disorder; MI, mild; MO, moderate; SE, severe impairment.

provided for the scoring were removed, and the VAS score was measured *a posteriori*. After the completion of the two evaluation sessions, raters were encouraged to share any comments on the evaluation sessions or on the CAPE-V$_f$ itself. Raters were blinded to the speaker's diagnosis, age, and sex.

Pilot testing

Before the launch of the study, our protocol was tested in a pilot trial conducted by four SLP students from the University of Toulouse, to verify the clarity of the instructions and to determine an average assessment time for each tool (CAPE-V$_f$ and GRBAS). Two students evaluated 19 voices with the CAPE-V$_f$, two others with the GRBAS. The estimated assessment times were 45 minutes for the CAPE-V$_f$ and 25 minutes for the GRBAS sessions (ie, about 1 minute and 30 seconds for each GRBAS rating, and 2 minutes and 35 seconds for each CAPE-V$_f$ rating). Minor modifications to the protocol were made based on the testers' feedback (eg, mentioning the possibility that some of the voices could be repeated within the listening list, as students had expressed surprise about hearing some samples twice).

Statistical analysis

Statistical analyses were performed using Jamovi version 2.5.3.[45] None of the CAPE-V$_f$ parameters passed the Shapiro-Wilk normality test (W = [0.76-0.91], $P < 0.001$); the GRBAS data were ordinal. Therefore, all subsequent analyses used nonparametric statistics.

Descriptive statistics

Descriptive statistics were computed to summarize the data's central tendency and variability. Although the CAPE-V$_f$ offers the possibility to evaluate the voice tasks separately, in the rare cases where this was done, the scores were averaged for statistical analysis. Second ratings of recordings that were presented twice for intra-rater reliability analysis were excluded from any other analysis.

Intra- and inter-rater reliability

For the CAPE-V$_f$ continuous VAS data, the intraclass correlation coefficient (ICC) was used, with a two-way random-effects model, treating our raters and voice recordings as random samples from their respective population; the ICC unit was "single rater" as the intended use of the CAPE-V is for one clinician to evaluate a voice; the relationship type was defined as "consistency" rather than "absolute agreement."

For CAPE-V$_f$ intra-rater reliability, the ICC was computed for each judge for each CAPE-V$_f$ parameter. The median of absolute differences was also computed to illustrate the extent of differences between repeated evaluations on the VAS. The repeated ratings of one rater were excluded for the CAPE-V$_f$ ratings, as they explained they went back to their first rating to provide a similar repeated value. Spearman correlations were computed between the years of experience and the ICCs for each vocal parameter to investigate a potential link between seniority and intra-rater reliability.

For CAPE-V$_f$ inter-rater reliability, the level of agreement between judges was assessed for each CAPE-V$_f$ parameter within each list. The median of absolute differences between the raters was also computed to illustrate the extent of differences between their ratings. An ICC < 0.5 indicated poor reliability; 0.5-0.75: moderate reliability; 0.75-0.9: good reliability; ICC > 0.90: excellent reliability.[46]

To assess the intra-rater reliability of the ordinal GRBAS ratings, Krippendorff's alpha was computed for each parameter and for each judge. For inter-rater reliability, Krippendorff's alpha was computed for each parameter and for each list. An alpha below 0.67 indicated poor agreement; 0.67-0.79: moderate agreement; 0.80-0.99: satisfactory agreement; 1: perfect agreement.[47]

Construct validity

To assess construct validity of the continuous CAPE-V$_f$ data in relation to the ordinal GRBAS ratings, Spearman correlations were computed between the equivalent vocal parameters in both tools: overall severity scores (CAPE-V$_f$) and G scores (GRBAS), as well as roughness, breathiness, and strain scores in both tools. Spearman correlations were further computed between the CAPE-V$_f$ overall severity and the VHI total score, as well as between the GRBAS G score and the VHI total score. A correlation coefficient between 0.0 and 0.10 indicated no correlation; 0.11-0.39: weak correlation; 0.40-0.69: moderate correlation; 0.70-0.89: strong correlation; ≥0.90: very strong correlation.[48]

Qualitative analysis on the use of the protocol

Comments made by the participants regarding the use of the CAPE-V$_f$ protocol were examined using thematic analysis. Observations on use patterns of the protocol were also made by MS while analyzing each individual form, eg, regarding the use of the additional blank VAS, the use of separate scoring for the three voice tasks, and the use of the "constant/intermittent" descriptors.

RESULTS

Descriptive statistics

Table 3 and Figure 1 describe the scores for each vocal parameter evaluated in the CAPE-V$_f$. The parameter with the lowest median score was pitch (6); the parameter with the highest median score was overall severity (28). The least variable parameter was breathiness, with an interquartile range (IQR) of 26; the highest variability was found for overall severity (IQR = 42.5).

The bar plot in Figure 2 illustrates the frequencies for the GRBAS parameter ratings. The sample contained 22% of nondysphonic ($G = 0$), 39% of mildly dysphonic ($G = 1$), 20% of moderately dysphonic ($G = 2$), and 19% of severely dysphonic ($G = 3$) ratings. Almost half of the ratings (49%) indicated no perceived breathiness ($B = 0$); 58% indicated no perceived asthenia. The ratings mostly showed perceived roughness (56% mild-to-moderate roughness, 11% severe roughness) and strain (51% mild-to-moderate strain, 12% severe strain).

Intra-rater reliability

The results for the CAPE-V$_f$ ratings (Table 4) indicated a good intra-rater reliability for overall severity, strain, and pitch (mean ICC: 0.83-0.89). Roughness, breathiness, and loudness presented moderate reliability (mean ICC: 0.57-0.73).

One rater (rater 1) was found to be very reliable in their evaluations; five raters demonstrated good reliability (raters 2, 3, 7, 9, and 10), and six showed moderate reliability (raters 4, 5, 6, 8, 11, and 12). The years of experience did not significantly correlate with the ICCs for any of the vocal parameters (r_s =0.07-0.40, $P = 0.28$-0.84).

TABLE 3.
Descriptive Data for Each CAPE-V$_f$ Parameter

	Overall severity	Roughness	Breathiness	Strain	Pitch	Loudness
N	155	155	155	155	129	127
Median	28.00	14.00	8.00	17.00	6.00	12.00
IQR	42.5	30	26	38.5	40	34.5
Minimum	0	0	0	0	0	0
Maximum	100.00	91.00	93.00	98.00	99.00	99.00

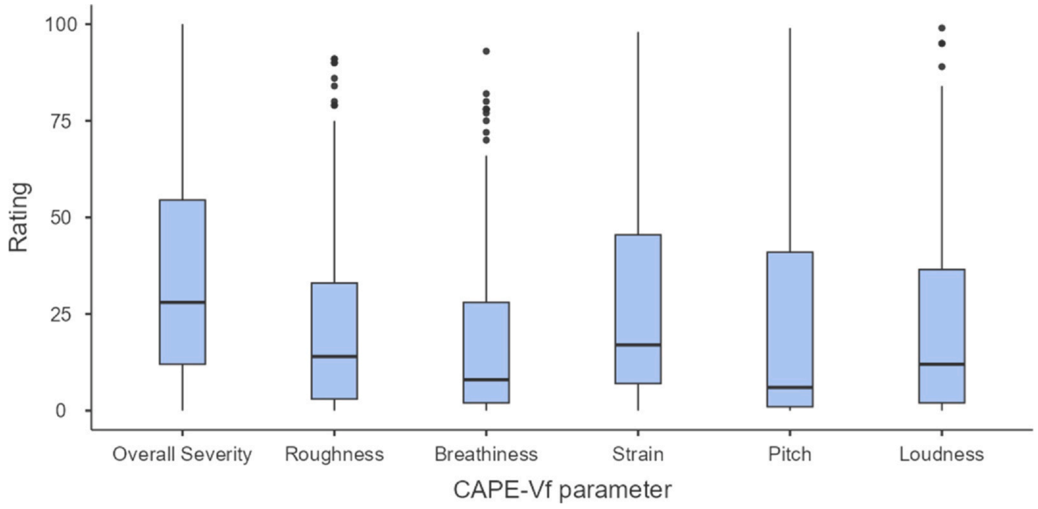

FIGURE 1. Boxplots for the ratings of each CAPE-V parameter.

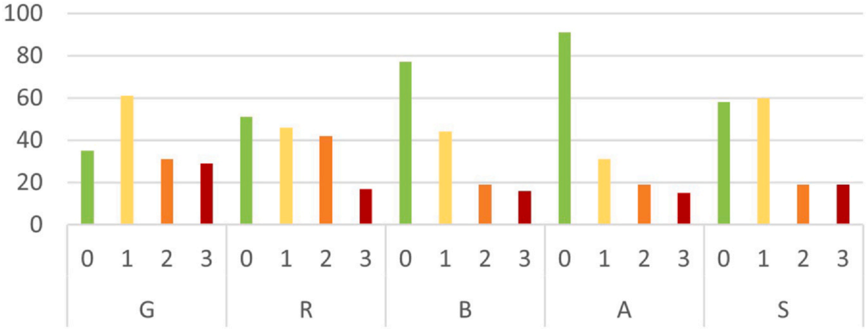

FIGURE 2. Bar plot for the frequencies (N) of severity ratings of the five GRBAS parameters.

TABLE 4.
Intra-Rater Reliability Results for Each Rater and Vocal Parameter of the CAPE-V

Rater	Exp. (years)	Overall severity *ICC (MAD)*	Roughness *ICC (MAD)*	Breathiness *ICC (MAD)*	Strain *ICC (MAD)*	Pitch *ICC (MAD)*	Loudness *ICC (MAD)*	Mean *ICC (MAD)*
1	10	0.94*** (11)	0.96*** (11)	0.82*** (15)	0.86** (11)	0.99*** (1)	0.83** (2.5)	**0.90 (9)**
2	6	0.92** (3)	0.69* (6)	*NA*	0.90** (6)	0.98*** (7)	0.97*** (5)	**0.89 (6)**
3	7	0.74* (10)	0.75* (4)	0.98*** (4)	0.88** (5)	0.85** (3)	0.82** (6)	**0.84 (5)**
4	6	0.90** (4)	0.92** (1)	0.87* (2)	0.70* (5)	0.65* (1)	0.35 (4)	**0.73 (3)**
5	20	0.81** (6)	0.80** (14)	0.37 (19)	0.85** (15)	*NA*	*NA*	**0.71 (14)**
6	8	0.70* (13)	0.84** (7)	0.15 (13)	0.81** (10)	0.86** (1)	0.5 (8)	**0.64 (10.5)**
7	5	0.92** (6)	0.63* (16)	0.42 (7)	0.83** (8)	0.99*** (1.5)	*NA*	**0.76 (7)**
8	15	0.91** (14)	0.81** (6)	0.19 (2)	0.69* (11)	0.99*** (2)	0.84** (8)	**0.74 (7)**
9	25	0.95*** (8)	0.97*** (4)	0.92** (5)	0.93*** (9)	0.64* (18)	0.87** (8)	**0.88 (8)**
10	1	0.99*** (3)	0.98*** (4)	0.60 (1)	0.93*** (3)	0.95*** (2)	0.41 (6)	**0.81 (2.5)**
11	23	0.92** (8)	0.16 (7)	0.77* (6)	0.98*** (6)	*NA*	*NA*	**0.71 (6.5)**
12	25	0.99*** (6)	0.22 (6)	0.15 (6)	0.57 (10)	0.89** (5)	0.77* (2)	**0.60 (6)**
Mean	12.6	**0.89 (7.7)**	**0.73 (7.2)**	**0.57 (7.3)**	**0.83 (8.3)**	**0.88 (4.2)**	**0.71 (5.5)**	**0.77 (7)**

Note: *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$; green = excellent reliability (ICC ≥ 0.90), blue = good reliability (ICC [0.75-0.90]), orange = moderate reliability (ICC [0.5-0.75]), red = low reliability (ICC < 0.5). *Abbreviations:* Exp., years of experience in voice therapy; MAD, median of absolute differences.

For the GRBAS (Table 5), the overall grade of dysphonia showed a satisfactory degree of agreement between repeated evaluations (mean $\alpha \geq 0.80$). Roughness, breathiness, and asthenia presented a moderate inter-rater agreement (mean α: [0.67-0.79]), while strain showed a low degree of agreement. The reliability of each rater averaged over all vocal parameters was satisfactory for four judges, moderate for four others, and low for the remaining five.

Inter-rater reliability

For the CAPE-V$_f$ (Table 6), the overall severity showed good inter-rater reliability (ICCm: [0.75-0.90]). Roughness, breathiness, strain, pitch, and loudness demonstrated overall low reliability (ICCm < 0.50). The median of absolute differences did not exceed 18.9 mm on the VAS for any of the CAPE-V$_f$ parameters.

All five GRBAS parameters (Table 7) showed low inter-rater reliability values (mean $\alpha < 0.67$).

TABLE 5.
Intra-Rater (Krippendorff's Alpha, α) Reliability Results for Each Rater and Each GRBAS Vocal Parameter

Rater	G	R	B	A	S	Mean α
1	0.63	0.88	0.86	0.75	0.82	**0.79**
2	1	0.39	0.52	*NA*	0.54	**0.61**
3	0.88	0.67	1	0.83	0.8	**0.84**
4	1	1	0.71	1	0.35	**0.81**
5	0.69	0.39	0.47	0.86	0.76	**0.63**
6	0.77	0.88	0.71	0.85	0.64	**0.77**
7	0.37	0.52	0.45	0.88	0.71	**0.59**
8	0.8	0.77	0.82	1	0.2	**0.72**
9	0.86	0.88	0.25	0.52	0.38	**0.58**
10	0.85	0.77	0.63	0.64	1	**0.78**
11	0.86	0.88	0.99	*NA*	0.49	**0.81**
12	0.84	0.88	1	0.84	0.77	**0.87**
13	0.85	0.51	0.67	0.33	0.78	**0.63**
Mean	**0.8**	**0.72**	**0.7**	**0.77**	**0.63**	**0.72**

Note: Green = perfect agreement (α:1), blue = satisfactory agreement (α: [0.80-0.99]), orange = moderate agreement (α: [0.67-0.79]), red = low reliability (α < 67).

Construct validity

Table 8 shows the Spearman correlation coefficients between the comparable parameters from the CAPE-V_f and the GRBAS. Given the low inter-rater reliability, the correlations were calculated for each rater rather than using mean scores across raters. A strong correlation (mean *r*: [0.70-0.89]) was measured between the CAPE-V_f overall severity and the G of the GRBAS, as well as a moderate correlation (mean *r*: [0.40-0.69]) between the roughness, the breathiness, and strain scores of the CAPE-V_f and the GRBAS.

The total VHI score was moderately correlated (mean *r*: [0.40-0.69]) with the overall severity scores of the CAPE-V_f and with the G of the GRBAS.

Qualitative analysis on the use of the protocol

Underutilized features

Some features of the protocol were scarcely used by the raters, starting with separate scoring of the three voice tasks. Out of 1366 evaluations, vocal tasks were evaluated separately only 104 times. When performing separate scorings, the raters showed a high precision in their distinct evaluations: the smallest difference between tasks for the same parameter was 5 mm. The two blank VAS were only filled out three times, by the same rater, who added the parameters “breathing,” “hard onset,” and “fry.” The constant/intermittent nature of each parameter was only evaluated by three judges: one used it almost systematically (100 times in total), one used it six times only, and the third one five times.

Pitch and loudness evaluation

Eight raters failed to systematically qualify the nature of the pitch and/or loudness impairment (too high/low, too soft/loud), despite quantifying them using the VAS. Quantitative ratings without qualifying the nature were observed 64 times out of 156: one rater did so once, five raters between 4 and 7 times, and two raters between 14 and 20 times. Conversely, some judges left a comment on the type of impairment without quantifying it on the VAS (total missing VAS scorings: 29/156 for pitch, 27/156 for loudness).

Use of comments

Most raters made use of comments on resonance and additional vocal characteristics, except for one rater who did not leave any additional comments. A total of 134 out of 228 possible comments were left for the 41 voices included in our study. Approximately 31% of the voices initially rated as “unimpaired” received a comment (including two positive comments on the speaker’s timbre). This number increased to 55% for slightly impaired voices, to 58% for moderately impaired voices, and to 85% for severely impaired voices.

DISCUSSION

The auditory-perceptual assessment conducted by SLPs is an essential component of the clinical voice assessment. Due to the lack of tools available in French, this study aimed to validate the European French version of the

TABLE 6.
Inter-Rater Reliability Results for the CAPE-V$_f$ Ratings, by List and by Vocal Parameter

List	Overall Severity *ICC (MAD)*	Roughness *ICC (MAD)*	Breathiness *ICC (MAD)*	Strain *ICC (MAD)*	Pitch *ICC (MAD)*	Loudness *ICC (MAD)*
1	0.75** (13)	0.67** (10.5)	0 (8)	0.59* (13)	0.96*** (5.5)	0.33 (8)
2	0.64*** (15.3)	0.26 (9)	0.77*** (9)	-0.15 (13)	0.05 (11.5)	0.18 (6)
3	0.82*** (14.3)	0.68*** (18)	0.48** (20.7)	0.17 (31)	0.16 (25.3)	-0.15 (16.5)
4	0.85*** (17.1)	0.28**(21.1)	0.53*** (18.8)	0.79***(18.6)	0.54*** (16.3)	0.66*** (21.2)
Mean	**0.77 (14.9)**	**0.47 (14.7)**	**0.45 (14.1)**	**0.35 (18.9)**	**0.43 (14.7)**	**0.26 (12.9)**

Note: *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$; green = excellent reliability (ICC ≥ 0.90), blue = good reliability (ICC [0.75-0.90]), orange = moderate reliability (ICC [0.5-0.75]), red = low reliability (ICC < 0.5). *Abbreviations:* MAD, median of absolute differences.

CAPE-V (CAPE-V$_f$) to expand the available tools for voice experts in France.

Reliability of the CAPE-V$_f$

Our study hypothesized that the CAPE-V$_f$ would demonstrate good intra- and inter-rater reliability (ICC ≥ 0.75). Reliability results in previous studies have shown significant variation, largely due to methodological differences, making comparisons difficult. This heterogeneity relates to factors such as the occupation of the raters (eg, inclusion of ENTs[28,30,32]) and their familiarity (eg,[33]) or unfamiliarity (eg,[32]) with the CAPE-V protocol; the use of a time gap between CAPE-V and GRBAS ratings (eg,[23,26]); the use of anchor voices to improve inter-rater reliability (eg,[31,33]); and the number of raters and voice samples. For a comprehensive review of methodologies, the reader is referred to.[36,38]

In this study, a notable strength was the use of 13 raters, which is higher than most previous CAPE-V adaptation studies (eg, two raters in,[26–28] three raters in,[25,31,33,35] four raters in,[30,34] and five raters in[29]). We chose to compare our reliability results to three studies that matched our methods more closely in terms of numbers of raters and voice samples: (1) the European Portuguese (EP) adaptation,[23] including 14 experienced SLPs, 10 dysphonic and 10 control voices, a 1-week time gap between CAPE-V and GRBAS ratings, without anchor voices, and using all three CAPE-V phonation tasks; (2) the Brazilian Portuguese (BP) adaptation,[22] including nine experienced SLPs, 10 euphonic, 10 mild, 10 moderate, and 10 severely dysphonic voices, a 48-72-hour time gap between CAPE-V and GRBAS ratings, with anchor voices, using the sustained vowel /a/ and the CAPE-V sentences; (3) the validity study for the American English (AE) version,[14] including 21 experienced SLPs, 13 mild, 11 moderate, and 13 severely dysphonic voices and 22 “normal” (*sic*) voices, a 48-72-hour gap between CAPE-V and GRBAS ratings, with anchor voices, using only conversational speech samples.

The intra-rater reliability in the present study met the 0.75 threshold for three of the six parameters and for half of the raters. Overall severity had the highest intra-rater reliability (mean ICC = 0.89), a general trend highlighted by Mahalingam et al[36] in their literature review. This parameter was followed closely by pitch (mean ICC = 0.88) and strain (mean ICC=0.83). Roughness and loudness showed moderate reliability (mean ICC = 0.73 and 0.71, respectively), while breathiness had the lowest intra-rater reliability (mean ICC = 0.57). The relatively low breathiness reliability could be due to the low level of breathiness in the voice samples (median score of 8 on the CAPE-V$_f$ and 49% of the voices with a score of 0 for the B parameter of the GRBAS) and the potential interaction between roughness and breathiness (67% of the voices presented with roughness), which may have led to a perceptual rating bias.[49,50] At the rater level, half of the SLPs showed good-to-excellent reliability in their scores; none of them showed a low reliability of repeated ratings. No clear trend of higher experience leading to higher reliability was observed. For example, rater 1, with 10 years of experience, had high ICC values for all characteristics; however, rater 12, with 25 years of experience, showed a relatively lower intra-rater reliability for several characteristics; similarly, rater 10, with just 1 year of experience, has a relatively high ICC for most characteristics, suggesting that more experience does not necessarily result in higher reliability. A direct comparison of our intra-rater results with the three studies cited above is challenging, as two of them (AE and EP) used Pearson’s correlations instead of ICCs, and BP did not

TABLE 7.
Inter-Rater (Krippendorff's Alpha, α) Reliability Results for Each List and Each GRBAS Vocal Parameter

List	G	R	B	A	S
1	*0.61*	*0.49*	*0.69*	*-0.34*	*0.08*
2	0.47	0.09	0.59	0.54	-0.11
3	0.62	0.37	0.49	0.46	-0.16
4	0.75	0.46	0.33	0.47	0.64
Mean α	**0.61**	**0.35**	**0.53**	**0.28**	**0.11**

Note: Green = perfect agreement (α: 1), blue = satisfactory agreement (α: [0.80-0.99]), orange = moderate agreement (α: [0.67-0.79]), red = low reliability (α < 67).

TABLE 8.
Spearman Correlations Between the Comparable Vocal Parameters of the CAPE-V$_f$ and the GRBAS per Rater, as Well as Between the VHI Total Score and the CAPE-V$_f$ Overall Severity and the GRBAS Grade, Respectively

Rater	Overall severity/G	Roughness/R	Breathiness/B	Strain/S	Overall severity-VHI$_{total}$	G-VHI$_{total}$
1	0.85***	0.95***	0.93***	0.81**	0.47	0.49
2	0.81**	0.68*	-0.25	0.68*	0.31	0.34
3	0.82**	0.41	0.67*	-0.11	0.69*	0.78**
4	0.73**	0.70*	0.78**	0.42	0.26	0.75**
5	0.91***	0.47	0.78**	0.51	0.84**	0.87***
6	0.94***	0.47	0.49	0.58*	0.81**	0.71**
7	0.89***	0.88***	0.63*	0.59*	0.71*	0.75**
8	0.72**	0.73**	0.75**	-0.28	0.69*	0.66*
9	0.84***	0.45	0.73**	0.74**	0.46	0.52
10	0.93***	0.85***	0.78**	0.70*	0.45	0.43
11	0.88***	0.58*	0.66*	0.60*	0.38	0.26
12	0.70*	0.62*	0.38	0.61*	0.43	0.37
13	0.94***	0.43	0.51	0.94***	0.39	0.42
Mean r_s	0.84	0.63	0.6	0.52	0.53	0.57

Note: r_s: Spearman correlation coefficient; *: P < 0.05; **: P < 0.01; ***: P < 0.001; green = very strong correlation (r_s: 0.9-1), blue = strong correlation (r_s: 0.70-89), orange = moderate correlation (r_s: 0.40-0.69), red = low reliability (r_s < 0.40).

specify the ICC model used. Nonetheless, the CAPE-V ICC values from the present study and from BP—both using ICCs—were largely consistent, except for a much higher intra-rater reliability for pitch in our study (ICC: 0.88 vs. −0.05 in BP). In both studies, the highest intra-rater reliability was achieved for overall severity (ICC: 0.89 and 0.86, respectively), followed closely by strain (ICC: 0.83 and 0.85, respectively). For the GRBAS ratings, BP used Cohen's Kappa, AE used Spearman's rho, and EP did not report the results. The low intra-rater reliability measured for the strain parameter of the GRBAS in the present study is consistent with the trends observed in AE and BP, as well as in a previous study specifically targeting the test-retest reliability of GRBAS ratings.[51] Finally, both the BP and the AE study found higher intra-rater reliability for the CAPE-V as compared with the GRBAS ratings, as in the present study. Overall, it can be concluded that repeated ratings using the CAPE-V$_f$ protocol are more reliable than GRBAS ratings.

The inter-rater reliability was generally low for most parameters, with only overall severity showing good reliability (mean ICC = 0.77). This suggests that while individual raters may have a similar perception of the overall severity of a voice disorder and are consistent in their own

TABLE 9.
Inter-Rater ICC Results for the CAPE-V Ratings in the Three Comparison Studies and in the Present Study

Study	Overall Severity	Roughness	Breathiness	Strain	Pitch	Loudness
BP	0.86	0.67	0.71	0.32	0.26	0.68
EP	0.96	0.92	0.95	0.84	0.86	0.90
AE	0.76	0.62	0.60	0.56	0.54	0.28
F	**0.77**	**0.47**	**0.45**	**0.35**	**0.43**	**0.26**

Abbreviations: BP, Brazilian Portuguese; EP, European Portuguese; AE, American English; F, French.

assessments (high intra-rater reliability), they may vary in their weighting of specific voice qualities. This is consistent with the well-reported observation that listeners differ in their perceptual strategies when rating isolated voice parameters,[52–54] questioning the existence of a common perceptual space.

All three comparison studies (AE, BP, and EP) also used ICCs for the investigation of inter-rater reliability of the CAPE-V, allowing for a more direct comparison of the results. Overall, the trend of higher ICCs for overall severity is clearly shown in all four studies (Table 9) as compared with lower inter-rater reliability for all other parameters. Strain and pitch showed lower inter-rater reliability both in the BP and in the AE validation studies, similar to our results. Loudness was also found to be unreliable in the AE validation study. In the present study, some ratings only included a qualitative evaluation of pitch and loudness using a verbal descriptor for the type of impairment (too soft/loud or too low/high), not using the VAS; other ratings consisted of the VAS, without specifying the nature of the alteration, as has been observed in Nagle's study about the clinical use of the CAPE-V scales[40]; hence, the inter-rater reliability results should be interpreted with caution. Furthermore, this observation might reflect a lack of practice in the quantitative perceptual rating of pitch and loudness, as these parameters are not included in the GRBAS, with which the raters are most familiar. This also begs the question of the clinical relevance of rating loudness and pitch quantitatively on a VAS, especially considering the ease of extraction of their acoustic correlates (speaking fundamental frequency and sound pressure level).[40] Finally, the reliability of loudness ratings also strongly depends on the quality and stability of the recording input, which ideally requires calibration of the equipment with a sound-level meter and is influenced by factors such as mouth-to-microphone distance and recording settings.

Discrepancies of inter-rater reliability results between studies might be explained to some degree by various methodological differences. EP found surprisingly high ICCs for the inter-rater reliability of all vocal parameters (ICC > 0.84). We hypothesize that this might be explained partly by the fact that their voice samples only included 20 speakers, of which 50% were control speakers without a voice impairment; yet, it has been demonstrated that listeners agree better on normal and on severely dysphonic voices, while the midrange of the severity continuum results in higher inter-rater variability.[55,56] In the present study, the voice sample was mostly composed of mild to moderately dysphonic voices, which might partly explain the lower inter-rater reliability results. Differences in the ICC model used might also account for some variance in the inter-rater reliability results: AE used two-way random-effects, single-rater ICCs (unclear if absolute agreement or consistency); EP used a two-way mixed-effects ICC model (unclear if single or multiple raters, agreement or consistency); BP did not specify the model used. This observation highlights the importance of providing sufficient methodological details when reporting study methods and results, to allow for replication studies and comparison of results. Another methodological factor affecting inter-rater reliability is the choice of the voice task/stimuli on which the ratings are performed, which has been extensively shown to impact both the vocal production and its perception.[55,57–59] In the present study, we used all three voice tasks of the CAPE-V_f (sustained vowel, sentences, and semispontaneous speech), similar to EP, while BP did not include the semispontaneous speech samples, and AE only used the latter. Finally, the lower inter-rater reliability results for the CAPE-V ratings observed in our results, could also partly be explained by the absence of severity markers (mild, moderate, and severe) under the VAS. This modification was made to avoid clustering of points under the descriptors, as highlighted in.[24,39,40] The absence of these markers under the VAS might thus have impacted the inter-rater reliability, as each rater referred to their own standards regarding the relationship between the numeric score on the VAS and the perceived severity of the impairment. This modification to the VAS allowed raters to evaluate the severity of the impairment on a true continuum, thereby restoring the advantages of this type of scale, including its superior sensitivity to changes. This trade-off between sensitivity and reliability has a clinical importance: for example, a voice that changes from a breathiness score of 40/100 to 20/100 would show a positive evolution that might not necessarily be noted if one relies on the categorical markers, as both scores are considered "moderate impairments."

Overall, given the good intra-rater reliability but lower inter-rater reliability, our first hypothesis is partially validated. It is important to consider the reliability results of the CAPE-V_f within the context of its clinical utility: in clinical settings, consistent intra-rater assessments are crucial for tracking patient progress over time. The lower inter-rater reliability is less concerning, as clinicians rarely rely solely on perceptual ratings for communication.

Construct validity of the CAPE-V_f

As in previous validation studies,[36] the CAPE-V_f demonstrated strong convergent validity, particularly in the correlation between its overall severity and the Grade ratings on the GRBAS (average $r_s = 0.84$). This suggests that the CAPE-V_f effectively captures the severity of voice disorders. However, moderate correlations between the CAPE-V_f and GRBAS ratings for roughness, breathiness, and strain (average $r_s = 0.63$, 0.60, and 0.52, respectively) indicate some divergence in how these tools assess specific voice qualities. These results are consistent with previous studies showing lower correlations for breathiness and strain.[36] Our hypothesis relating to the CAPE-V_f's convergent validity is thus partially validated.

The discriminant validity of the CAPE-V_f was also confirmed, as the correlation between overall severity on the CAPE-V_f and the total VHI scores was weaker (average $r_s = 0.53$). Considering the high correlation of the GRBAS Grade with the CAPE-V_f overall severity, it is not surprising that its correlation with the total VHI score is similar as well (average $r_s = 0.57$). This result confirms that while the VHI is a patient-reported outcome measure reflecting the voice-related functional impact in the patient's everyday life—influenced by various personal, contextual, and sociological factors—both the GRBAS and the CAPE-V_f as clinician-rated tools target a different aspect of the multidimensional voice assessment. They allow to qualify and quantify the auditory-perceptual aspects of the patient's vocal function and contribute to hypotheses regarding the anatomophysiological mechanisms of the voice disorder. Both sources of information are imperative to understand the voice disorder in the diversity of its manifestations.

Qualitative analysis

The qualitative analysis of the CAPE-V_f forms and the feedback from the students and raters who participated in the study highlighted areas for potential improvement in the CAPE-V_f.

The confusion caused by the evaluation of pitch and loudness highlights the need to either clarify instructions or revise the scales used. This issue was already raised during the development of the French adaptation of the CAPE-V.[39] Nagle[40] also observed that some clinicians failed to use these features, potentially because they were uncomfortable rating loudness on recordings or prefer relying on their acoustic correlates. To improve the ratings on the CAPE-V form, we envision two options for revised scales for the assessment of pitch and loudness: either a simple qualitative verbal description of the alteration (without a VAS), or a bipolar VAS labeled at each extremity of the scale, with "normality" as the central point (eg, for loudness: "too soft–normal–too loud").

The underutilization of certain features suggests the need for better integration or clearer instructions in the CAPE-V_f protocol. The infrequent use of the additional blank VAS has been reported in previous research.[40] Nevertheless, their inclusion—even if unused by some clinicians—does not affect the use of the tool; we therefore advocate retaining this option that could prove important in specific clinical contexts or patient groups (eg, in neurological or pediatric voice disorders with unique characteristics such as tremor, voice breaks, or inconsistent pitch control), allowing for a tailored rating experience. The same applies to the possibility to rate tasks separately on the VAS: the high precision in separate evaluations despite their infrequent use suggests that when raters are confident in using these features, the resulting information may prove useful both in clinical and in research use cases. Meanwhile, the rare use of the "constant/intermittent" descriptors of the nature of each parameter could reflect a lack of voice samples representative of intermittent vocal alterations in our study and does not in itself question the relevance of this feature.

The extensive use of comments for resonance and additional vocal characteristics highlights a significant advantage of the CAPE-V over the GRBAS. The ability to provide qualitative descriptions alongside quantitative ratings enhances the tool's clinical utility, allowing for a clearer understanding of the rater's intentions, and partially mitigates the low inter-rater reliability of its quantitative results.

Finally, the longer administration time for the CAPE-V_f, compared with the GRBAS, is a consideration for its user. Factors that may have contributed to this difference include the raters' unfamiliarity with the CAPE-V_f, the digital modality requiring more manipulations to fill out the protocol as compared with the GRBAS tables, and the greater number of vocal parameters evaluated. Clinicians and researchers must therefore consider the trade-off between time efficiency and the sensitivity and reliability of the perceptual assessment when choosing the appropriate auditory-perceptual assessment tool.

LIMITS AND FUTURE PERSPECTIVES

This study presents some limitations, mainly related to the participants and the perceptual rating sessions.

All participants were native French speakers from France, which limits the generalizability of the results, particularly considering the diversity of accents encountered in French-speaking voice clinics, in Europe, and worldwide. Sociocultural factors, including both the speaker's and the rater's language and regional accents, have been shown to impact the perception and description of voice quality,[15–21] especially in mildly impaired voices.[21] Hence, future studies should extend the validation of the CAPE-V_f to other Francophone populations, allowing for a broader and more culturally and linguistically adapted use of the CAPE-V_f. Validation with other voice clinicians, such as ENT specialists and phoniatricians, as well as with researchers, would also be desirable to confirm the relevance of using the CAPE-V_f in various clinical and research contexts. Of note, a separate version of the CAPE-V is currently being validated in Quebec French.

A potential bias was introduced by some participants mentioning their SLP treatment or underlying pathology during the semispontaneous speech task, despite our intention to mask the diagnosis. In clinical use cases, however, the diagnosis is most often known to the clinician and can impact their perceptual assessment. Similarly, for pitch ratings, the speaker's age and sex are known to the rater in clinical use cases. This information was not presented to the raters in the present study, forcing them to provide a decontextualized rating. Additionally, assessing voices solely based on audio recordings excluded the observation of physical manifestations of dysphonia (eg, posture, muscle tension, and breathing patterns), which are important in a comprehensive voice assessment and can enhance the perceptual assessments, particularly for the "strain" and "asthenia" parameters.[60] It can be assumed that real-world, multimodal evaluation in a clinical setting would therefore result in an even better reliability of the CAPE-V_f.

The imbalance in the composition of listening lists due to the dropout of eight of the original 21 judges also limits the robustness of the reliability and correlation results. Future studies should ensure a more balanced distribution of raters across different voice samples.

Further future perspectives emerge from the present study to continue improving the CAPE-V_f protocol, to optimize its use in clinical and research settings, and to expand our knowledge of vocal perception.

We made the deliberate choice not to offer training sessions or anchor voices in the present study to provide assessment conditions that match the intended use of the CAPE-V_f protocol in clinical settings. This choice reflects the methodologies of other CAPE-V adaptation studies such as.[24] However, other studies used anchor voices or provided a specific training (eg,[14,28,30,31,33]), or only included listeners that were already familiar with the CAPE-V protocol (eg,[23,26,28]). For a better inter-rater reliability of the CAPE-V_f, providing clinicians with anchor samples and training sessions is an important future perspective. Indeed, it is recognized in the literature that these strategies improve the degree of agreement and reduce score variability between raters by harmonizing the raters' internal referents of pathological voice qualities[61,62]. This training could be included in a future digitized version of the CAPE-V_f.

Despite the low inter-rater reliability, the median absolute differences did not exceed 18.9 mm on the VAS for any CAPE-V_f parameter. This suggests that, while raters may differ in their scoring, the overall clinical impact may be limited. Future studies should explore the minimal detectable change and minimal important difference to understand clinically meaningful changes on the scales and to the determine if the VAS scales are to be interpreted linearly.

Moreover, future studies could explicitly require the raters to assess each task separately, as in.[26] While this would not systematically be feasible in clinical contexts because of the time-consuming nature, it could inform about potential differences in the perception of vocal parameters depending on the task as well as about the importance of each task in the clinical voice assessment, adding to the fundamental literature about voice perception mechanisms. Specific methods such as the use of a free sorting task,[63–67] should also shed light on the underlying cognitive and perceptual mechanisms used to classify voices based on individual qualities and test the hypothesis that expert listeners are able to consistently isolate and focus their attention on individual voice parameters in complex voice samples.

Validation of the CAPE-V_f against instrumental measures could also be considered for future studies, to determine how physiological modifications, as evidenced by acoustic or aerodynamic measures, are reflected in the auditory perception of vocal parameters. However, it needs to be kept in mind that to date, overall severity, strain, breathiness, and roughness do not have consensual acoustic correlates, unlike pitch and loudness, which are easily measurable acoustically.[68] Similarly, it would be interesting to determine whether the sentences included in the CAPE-V_f protocol genuinely elicit the various vocal phenomena and behaviors they are meant to provoke. To this effect, a targeted perceptual evaluation of vocal behaviors (eg, presence/absence of hard vocal onsets) could be combined with visual analysis of spectrograms, as well as acoustic measures (eg, nasality for sentence 5, which could be quantified by the low tone-high tone ratio[69]) or aerodynamic measures (eg, subglottic pressure for sentence 4). This perspective seems particularly interesting to investigate as some studies[70,71] have revealed significant correlations between the CAPE-V and vocal acoustic and aerodynamic measures.

Finally, in light of the qualitative observations made in the present study, it also seems relevant to consider some modifications to the current CAPE-V_f protocol to improve its specificity and the efficiency of its use. Among these, we suggested modifying the scales for the pitch and loudness parameters. Considering the higher administration time for the CAPE-V_f, including the time needed to measure the VAS scoring, replacing the VAS by equal-appearing interval scales (EAI) could also be considered. Indeed, both types of scales have been shown to provide linearly related ratings of vocal roughness and breathiness.[72] The use of EAI instead of VAS has been implemented in the Mandarin[30] and in the Malay[25] CAPE-V adaptations, to increase the speed of ratings. Another possible solution to decrease the administration time could be to develop a digital version of the CAPE-V_f, in which case the rater could place a cursor on the VAS, allowing for automatic extraction of the measurement.

CONCLUSION

The CAPE-V$_f$ demonstrated good intra-rater reliability and construct validity, but lower inter-rater reliability for all vocal parameters but overall severity. The CAPE-V$_f$ is a comprehensive, standardized protocol, including sustained phonation, targeted evaluation of vocal behaviors through specifically designed sentences, as well as a more ecological semispontaneous speech task. The psychometric superiority of the CAPE-V over the GRBAS found in our results has been reported as a common tendency throughout the CAPE-V validation studies and is one of the arguments in favor of its clinical use for more robust voice assessments. The creation of a future digital version of the CAPE-V$_f$ could potentially address some of the shortcomings identified in the present study.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors extend their sincere appreciation to the speech-language pathologists who participated in the study for their valuable time and expertise. Special thanks go to the patients whose voices made this research possible.

References

1. Stachler RJ, Francis DO, Schwartz SR, et al. Clinical practice guideline: hoarseness (dysphonia) (update). *Otolaryngol Head Neck Surg.* 2018;158(1_suppl):S1–S42. https://doi.org/10.1177/0194599817751030.
2. Lechien JR, Geneid A, Bohlender JE, et al. Consensus for voice quality assessment in clinical practice: guidelines of the European Laryngological Society and Union of the European Phoniatricians. *Eur Arch Oto-Rhino-Laryngol.* 2023;280:5459–5473. https://doi.org/10.1007/s00405-023-08211-6.Published online September 14.
3. Behrman A. Common practices of voice therapists in the evaluation of patients. *J Voice.* 2005;19:454–469. https://doi.org/10.1016/j.jvoice.2004.08.004.
4. Oates J. Auditory-perceptual evaluation of disordered voice quality. *Folia Phoniatrica et Logopaedica.* 2009;61:49–56. https://doi.org/10.1159/000200768.
5. Barsties B, De Bodt M. Assessment of voice quality: current state-of-the-art. *Auris Nasus Larynx.* 2015;42:183–188. https://doi.org/10.1016/j.anl.2014.11.001.
6. Hirano M. *Clinical Examination of Voice.* Berlin, Germany: Springer Verlag; 1981.
7. Kempster GB, Gerratt BR, Verdolini Abbott K, et al. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol.* 2009;18:124–132. https://doi.org/10.1044/1058-0360(2008/08-0017).
8. Nemr K, Simões-Zenari M, Cordeiro GF, et al. GRBAS and Cape-V scales: high reliability and consensus when applied at different times. *J Voice.* 2012;26:812.e17–812.e22. https://doi.org/10.1016/j.jvoice.2012.03.005.
9. Giovanni A, de Saint-Victor S. Bilan clinique de la voix. *EMC Oto-Rhino-Laryngologie.* 2013;8:1–15.
10. Dejonckere PH, Obbens C, de Moor GM, Wieneke GH. Perceptual evaluation of dysphonia: Reliability and relevance. *Folia Phoniatrica et Logopaedica.* 1993;45:76–83. https://doi.org/10.1159/000266220.
11. Ghio A, Revis J, Smithson-Barrière D, et al. Reliability and correlations between overall severity, roughness and breathiness in the perception of dysphonic voices: investigating cognitive aspects. *J Voice.* 2021;38:136–143. https://doi.org/10.1016/j.jvoice.2021.07.010.
12. Karnell MP, Melton SD, Childes JM, et al. Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *J Voice.* 2007;21:576–590. https://doi.org/10.1016/j.jvoice.2006.05.001.
13. Fujiki RB, Thibeault SL. The relationship between auditory-perceptual rating scales and objective voice measures in children with voice disorders. *Am J Speech Lang Pathol.* 2021;30:228–238. https://doi.org/10.1044/2020_AJSLP-20-00188.
14. Zraick RI, Kempster GB, Connor NP, et al. Establishing validity of the consensus auditory-perceptual evaluation of voice (CAPE-V). *Am J Speech Lang Pathol.* 2011;20:14–22. https://doi.org/10.1044/1058-0360(2010/09-0105).
15. Dejonckere PH, Bradley P, Clemente P, et al. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. *Eur Arch Oto-Rhino-Laryngol.* 2001;258:77–82. https://doi.org/10.1007/s004050000299.
16. Yiu EML, Murdoch B, Hird K, et al. Cultural and language differences in voice quality perception: a preliminary investigation using synthesized signals. *Folia Phoniatrica et Logopaedica.* 2008;60:107–119. https://doi.org/10.1159/000119746.
17. Ghio A, Cantarella G, Weisz F, et al. Is the perception of dysphonia severity language-dependent? A comparison of French and Italian voice assessments. *Logoped Phoniatr Vocol.* 2015;40(1):36–43. https://doi.org/10.3109/14015439.2013.837503.
18. Yamaguchi H, Shrivastav R, Andrews ML, Niimi S. A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. *Folia Phoniatrica et Logopaedica.* 2003;55:147–157. https://doi.org/10.1159/000070726.
19. Järvinen K, Laukkanen AM, Geneid A. Voice quality in native and foreign languages investigated by inverse filtering and perceptual analyses. *J Voice.* 2017;31:261.e25–261.e31. https://doi.org/10.1016/j.jvoice.2016.05.003.
20. Engelbert A. Cross-linguistic effects on voice quality: a study on Brazilians' production of Portuguese and English Concordia Working Papers in Applied Linguistics; 2014;5:157–170.
21. Ghio A, Gasquet-Cyrus M, Roquel J, Giovanni, A. Perceptual interference between regional accent and voice/speech disorders In: Interspeech 2013 2013; ISCA, 2138-2142. doi: 10.21437/Interspeech.2013-506.
22. Behlau M, Rocha B, Englert M, Madazio G. Validation of the Brazilian Portuguese CAPE-V instrument—Br CAPE-V for auditory-perceptual analysis. *J Voice.* 2022;36:586.e15–586.e20. https://doi.org/10.1016/j.jvoice.2020.07.007.
23. de Almeida SC, Mendes AP, Kempster GB. The consensus auditory-perceptual evaluation of voice (CAPE-V) psychometric characteristics: II European Portuguese Version (II EP CAPE-V). *J Voice.* 2019;33:582.e5–582.e13. https://doi.org/10.1016/j.jvoice.2018.02.013.
24. Calaf N, Garcia-Quintana D. Development and validation of the bilingual Catalan/Spanish cross-cultural adaptation of the consensus auditory-perceptual evaluation of voice. *J Speech Lang Hear Res.* 2024;67:1072–1089. https://doi.org/10.1044/2024_JSLHR-23-00536.
25. Mohd Mossadeq N, Mohd Khairuddin KA, Zakaria MN. Cross-cultural adaptation of the consensus auditory-perceptual evaluation of voice (CAPE-V) into malay: a validity study. *J Voice.* 2022. https://doi.org/10.1016/j.jvoice.2022.05.018. S0892-1997(22):00151-5.
26. Núñez-Batalla F, Morato-Galán M, García-López I, Ávila-Menéndez A. Validation of the Spanish adaptation of the consensus auditory-perceptual evaluation of voice (CAPE-V). *Acta Otorrinolaringologica*

(English Edition). 2015;66:249–257. https://doi.org/10.1016/j.otoeng.2015.08.001.

27. Joshi A, Baheti I, Angadi V. Cultural and Linguistic adaptation of the consensus auditory-perceptual evaluation of voice (CAPE-V) Into Hindi. *J Speech Lang Hear Res.* 2020;63:3974–3981. https://doi.org/10.1044/2020_JSLHR-20-00348.
28. Ertan-Schlüter E, Demirhan E, Ünsal EM, Tadıhan-Özkan E. The Turkish version of the consensus auditory-perceptual evaluation of voice (CAPE-V): a reliability and validity study. *J Voice.* 2020;34:965.e13–965.e22. https://doi.org/10.1016/j.jvoice.2019.05.014.
29. Kondo K, Mizuta M, Kawai Y, et al. Development and validation of the Japanese version of the consensus auditory-perceptual evaluation of voice. *J Speech Lang Hear Res.* 2021;64:4754–4761. https://doi.org/10.1044/2021_JSLHR-21-00269.
30. Chen Z, Fang R, Zhang Y, et al. The Mandarin version of the consensus auditory-perceptual evaluation of voice (CAPE-V) and its reliability. *J Speech Lang Hearing Res.* 2018;61:2451–2457. https://doi.org/10.1044/2018_JSLHR-S-17-0386.
31. Mozzanica F, Ginocchio D, Borghi E, et al. Reliability and validity of the Italian version of the consensus auditory-perceptual evaluation of voice (CAPE-V). *Folia Phoniatrica et Logopaedica.* 2013;65:257–265. https://doi.org/10.1159/000356479.
32. Salary MN, Khoddami S, Drinnan M, et al. Validity and rater reliability of Persian version of the consensus auditory perceptual evaluation of voice. *Aud Vestib Res.* 2017;23:65–74.
33. Gunjawate DR, Ravi R, Bhagavan S. Reliability and validity of the Kannada version of the consensus auditory-perceptual evaluation of voice. *J Speech Lang Hear Res.* 2020;63:385–392. https://doi.org/10.1044/2019_JSLHR-19-00020.
34. Özcebe E, Aydinli FE, Tiğrak TK, et al. Reliability and validity of the Turkish version of the consensus auditory-perceptual evaluation of voice (CAPE-V). *J Voice.* 2019;33:382.e1–382.e10. https://doi.org/10.1016/j.jvoice.2017.11.013.
35. Venkatraman Y, Mahalingam S, Boominathan P. Development and validation of sentences in Tamil for psychoacoustic evaluation of voice using the consensus auditory-perceptual evaluation of voice. *J Speech Lang Hear Res.* 2022;65:4539–4556. https://doi.org/10.1044/2022_JSLHR-22-00169.
36. Mahalingam S, Venkatraman Y, Boominathan P. Cross-cultural adaptation and validation of consensus auditory perceptual evaluation of voice (CAPE-V): a systematic review. *J Voice.* 2024;38:630–640. https://doi.org/10.1016/j.jvoice.2021.10.022.
37. Dabirmoghaddam P, Khoramshahi H, Dehqan A, et al. Construct and discriminant validity of the Persian version of the consensus auditory perceptual evaluation of voice (CAPE-V). *J Voice.* 2022;36:876.e9–876.e15. https://doi.org/10.1016/j.jvoice.2020.09.023.
38. Narea-Veas MS, Farías PG, Vázquez Fernández P. Consensus auditory perceptual evaluation of voice (CAPE-V): revisión sistemática de los métodos utilizados para su adaptación y validación. *Revista de Investigación e Innovación en Ciencias de la Salud.* 2023;5:178–204. https://doi.org/10.46634/riics.206.
39. Pommée T, Mbagira D, Morsomme D. French-Language adaptation of the consensus auditory-perceptual evaluation of voice (CAPE-V). (Published online). *J Voice.* 2024. https://doi.org/10.1016/j.jvoice.2024.03.011.
40. Nagle KF. Clinical use of the CAPE-V scales: agreement, reliability and notes on voice quality. *J Voice.* 2022. https://doi.org/10.1016/j.jvoice.2022.11.014.S0892-1997(22):00366-6.
41. Boateng GO, Neilands TB, Frongillo EA, et al. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Front Public Health.* 2018;6:149. https://doi.org/10.3389/fpubh.2018.00149.
42. Piron A. OSTEOVOX, formation en thérapies manuelles et réhabilitation sensori motrice dédiée à la sphère oro-faciale. Published online 1998.
43. Jacobson BH, Johnson A, Grywalski C, et al. The Voice Handicap Index (VHI). *Am J Speech Lang Pathol.* 1997;6:66–70. https://doi.org/10.1044/1058-0360.0603.66.
44. Woisard V, Bodin S, Puech M. The Voice Handicap Index: impact of the translation in French on the validation. *Rev Laryngol Otol Rhinol (Bord).* 2004;125:307–312.
45. The jamovi project. jamovi. Published online 2024.
46. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016;15:155–163. https://doi.org/10.1016/j.jcm.2016.02.012.
47. Marzi G, Balzano M, Marchiori D. K-alpha calculator–Krippendorff's alpha calculator: a user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient. *MethodsX.* 2024;12:102545. https://doi.org/10.1016/j.mex.2023.102545.
48. Schober P, Boer C, Schwarte LA. Correlation coefficients. *Anesth Analg.* 2018;126:1763–1768. https://doi.org/10.1213/ANE.0000000000002864.
49. Park Y, Anand S, Kopf LM, et al. Interactions between breathy and rough voice qualities and their contributions to overall dysphonia severity. *J Speech Lang Hear Res.* 2022;65:4071–4084. https://doi.org/10.1044/2022_JSLHR-22-00012.
50. Millet B, Dejonckere PH. What determines the differences in perceptual rating of dysphonia between experienced raters? *Folia Phoniatrica et Logopaedica.* 1998;50:305–310. https://doi.org/10.1159/000021472.
51. De Bodt MS, Wuyts FL, Van de Heyning PH, Croux C. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice.* 1997;11:74–80. https://doi.org/10.1016/S0892-1997(97)80026-4.
52. Kreiman J, Gerratt BR. The perceptual structure of pathologic voice quality. *J Acoust Soc Am.* 1996;100:1787–1795. https://doi.org/10.1121/1.416074.
53. Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. *J Acoust Soc Am.* 2007;122:2354–2364. https://doi.org/10.1121/1.2770547.
54. Kreiman J, Gerratt BR. Sources of listener disagreement in voice quality assessment. *J Acoust Soc Am.* 2000;108:1867–1876. https://doi.org/10.1121/1.1289362.
55. Kreiman J, Gerratt BR, Kempster GB, et al. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res.* 1993;36:21–40. https://doi.org/10.1044/jshr.3601.21.
56. Kreiman J, Gerratt BR. Validity of rating scale measures of voice quality. *J Acoust Soc Am.* 1998;104:1598–1608. https://doi.org/10.1121/1.424372.
57. Barsties B, Maryn Y. External validation of the acoustic voice quality index Version 03.01 with extended representativity. *Ann Otol Rhinol Laryngol.* 2016;125:571–583. https://doi.org/10.1177/0003489416636131.
58. Wolfe V, Cornell R, Fitch J. Sentence/vowel correlation in the evaluation of dysphonia. *J Voice.* 1995;9:297–303. https://doi.org/10.1016/S0892-1997(05)80237-1.
59. Gerratt BR, Kreiman J, Garellek M. Comparing measures of voice quality from sustained phonation and continuous speech. *J Speech Lang Hear Res.* 2016;59:994–1001. https://doi.org/10.1044/2016_JSLHR-S-15-0307.
60. Payten CL, Chiapello G, Weir KA, Madill CJ. Frameworks, terminology and definitions used for the classification of voice disorders: a scoping review. *J Voice.* 2024;38:1070–1087. https://doi.org/10.1016/j.jvoice.2022.02.009.
61. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced Listeners' judgments of dysphonic voice. *J Voice.* 2006;20:527–544. https://doi.org/10.1016/j.jvoice.2005.08.007.
62. Ghio A, Dufour S, Rouaze M, et al. Mise au point et évaluation d'un protocole d'apprentissage de jugement perceptif de la sévérité de dysphonies sur de la parole naturelle. *Revue de Laryngologie-Otologie-Rhinologie.* 2011;132:19–27.
63. Woisard V, Gaillard P, Duez D. Free sorting task of speech disorders by expert and non expert listeners. *Rev Laryngol Otol Rhinol (Bord).* 2012;133:9–17.
64. Lansford KL, Liss JM, Norton RE. Free-classification of perceptually similar speakers with dysarthria. *J Speech Lang Hear Res.* 2014;57:2051–2064. https://doi.org/10.1044/2014_JSLHR-S-13-0177.

65. Clopper CG, Pisoni DB. Free classification of regional dialects of American English. *J Phon.* 2007;35:421–438. https://doi.org/10.1016/j.wocn.2006.06.001.
66. Morange S, Dubois D, Fontaine JM. Perception of recorded singing voice quality and expertise: cognitive linguistics and acoustic approaches. *J Voice.* 2010;24:450–457. https://doi.org/10.1016/j.jvoice.2008.08.006.
67. Clopper CG. Auditory free classification: methods and analysis. *Behav Res Methods.* 2008;40:575–581. https://doi.org/10.3758/BRM.40.2.575.
68. Nagle KF. Emerging scientist: challenges to CAPE-V as a standard. *Perspect ASHA Spec Interest Groups.* 2016;1:47–53. https://doi.org/10.1044/persp1.SIG3.47.
69. Lee GS, Wang CP, Yang CCH, Kuo TBJ. Voice low tone to high tone ratio: a potential quantitative index for vowel [a:] and its nasalization. *IEEE Trans Biomed Eng.* 2006;53:1437–1439. https://doi.org/10.1109/TBME.2006.873694.
70. EL-Adawy A, El-Rabie Ahmed M, Hassan E, Rezk Mohammed I. Modified GRBAS versus Cape-v scale for assessment of voice quality: correlation with acoustic and aerodynamics measurement for Arabic speaking subjects. *Sci Med J.* 2011;23:1–14.
71. Arabi A, Tarameshlu M, Behroozmand R, Ghelichi L. Correlation between auditory-perceptual parameters and acoustic characteristics of voice in theater actors. *Middle East J Rehabil Health Studies.* 2023;10:e131241. https://doi.org/10.5812/mejrh-131241.
72. Yiu EM -L, Ng C. Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness. *Clin Linguist Phon.* 2004;18:211–229. https://doi.org/10.1080/0269920042000193599.